Forensic Science International 343 (2023) 111543

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Not all identification conclusions are equal: Quantifying the strength of fingerprint decisions

Thomas Busey [a,*], Meredith Coon [b]

[a] Indiana University, USA
[b] Baltimore Police Department, USA

ARTICLE INFO

Article history:
Received 1 August 2022
Received in revised form 20 December 2022
Accepted 21 December 2022
Available online 2 January 2023

Keywords:
Likelihood ratio
Strength of evidence
Categorical conclusions
Fingerprints
Pattern comparison

ABSTRACT

In the pattern comparison disciplines, forensic practitioners evaluate two impressions with respect to the same-source and different-sources propositions. The results are communicated using a pre-determined conclusion scale, and in the friction ridge discipline Identification is typically the highest category on the scale for reporting evidence supporting the same source proposition. Although error rates have been measured in most disciplines, there are no widespread quantitative approaches and therefore most conclusions rely on subjective human evaluations. The current work uses articulation decisions provided by fingerprint examiners in error rate studies to produce a quantitative likelihood ratio measure that characterizes the strength of the support for the two propositions. We use an ordered probit model to summarize the distribution of responses of examiners who participated in error rate and validation studies. We then aggregate the data for all image pairs in a database to construct a set of likelihood ratios based on the ratio of the two strength-of-support values. We find that these values are modest relative to values typically produced by DNA analysis or implied by current fingerprint articulation language. The technique can be applied to any pattern comparison discipline for which error-rate data is available, and therefore can be used to appropriately weigh the evidence from different disciplines.

Forensic practitioners in the pattern comparison disciplines such as fingerprints, toolmarks, firearms, footwear, or handwriting typically compare an unknown impression or sample against an impression or sample from a known source. The results are communicated either using a categorical conclusion scale or a subjective likelihood ratio, and in most cases these are conducted by human examiners rather than by computer-based measures of similarity. In the latent print discipline, print impressions that are collected from crime scenes are compared with known impressions either taken from a suspect or retrieved from a computer database. Variations in deposition pressure, contact, and surface material produce differences in appearance for impressions of the same skin, and large database searches introduce the possibility of close nonmatches that make some comparisons challenging. In the US, fingerprint examiners have traditionally expressed their conclusions using one of three different articulation statements: Identification, meaning they believe that the two impressions came from the same finger; Exclusion, meaning they believe that the two impressions came from different fingers; or Inconclusive, meaning they cannot make any other determination. A latent print may also be judged not of value, in which case it typically is not compared. Identification decisions are usually verified by a second examiner, although this practice is not universal.

Articulation statements such as Identification or Exclusion are posteriors, in the sense that they are statements about a hypothesis (i.e. the same finger made the two impressions) rather than a statement about probability of observing a particular degree of correspondence given the same-source and different-sources propositions. Because the statements are posteriors, they transpose the conditional [29], subsume the role of factfinder that is typically the domain of the jury or judge, and can lead to "ipse dixit" (because I said so) reasoning. For these reasons, the National Research Council criticized the lack of scientific foundation for these decisions [1].

An alternative approach is to characterize the strength of the evidence in terms of a *likelihood ratio*. One definition is this:

> *"A likelihood ratio compares the probabilities of observing the evidence under two different hypotheses."* ([4], *p. 2)*

* Correspondence to: Department of Psychological and Brain Sciences, Indiana University, 1101 E 10th St., Bloomington, IN 47408, USA.
E-mail address: busey@indiana.edu (T. Busey).

https://doi.org/10.1016/j.forsciint.2022.111543
0379-0738/

In an expanded form, a definition might look like this, which is our summary of several different definitions and articulation documents:

> A likelihood ratio expresses a forensic examiner's assessment of the relative probabilities of the observations if one proposition is true versus if the other proposition is true. The two propositions typically come from the case, and may involve the proposition that the suspect contributed to the sample vs some unknown individual contributed to the sample.

One way that the finder of fact can use the likelihood ratio is through Bayesian updating. They first evaluate how much more guilty than innocent they believe the suspect is before hearing the forensic evidence (which represents the prior odds), and then multiply that ratio by the likelihood ratio provided by the forensic evidence to produce an updated posterior ratio that represents how much more guilty than innocent the suspect appears after hearing the new forensic evidence. This latter value represents the posterior odds, which can then become the prior odds when the next likelihood ratio is introduced during the case. This process assumes that the evidence is both probative and viewed as credible.

The likelihood ratio metric is widely viewed as a straightforward way to integrate new information with existing evidence [4,9]. There are challenges for novices in interpreting likelihood ratios [24], but the logical foundations for the likelihood ratio are widely accepted [2]. In Europe, the benefits of likelihood ratios have been acknowledged through explicit policy recommendations, and in the absence of quantitative approaches European practitioners have adopted a subjective likelihood ratio [2].

There have been some efforts to derive a quantitative likelihood ratio-type measure for fingerprints. Minutiae-based likelihood ratios based on a limited number of minutiae and assumptions about within-sample variability have been proposed [16,17,15] but these have not seen widespread adoption and are specific to fingerprints. Swofford et al. [28] developed a method that uses examiner-annotated features (placed on minutiae on the impressions) to produce a similarity statistic that is evaluated against a data set to create the probability of obtaining the degree of observed correspondence from mated and nonmated impressions. Although this method is used in the Defense Forensic Science Center/USACIL [25], the articulation language is thought to be confusing to both examiners and novices [5] and this method is not currently widely adopted by the community. The model computes statistics on a limited number of features, and only provides values in scenarios where correspondence is thought to be observed. The model also relies on cumulative probabilities and therefore is not a true likelihood ratio. Barriers to adoption may include the fact that government agencies are reluctant to take on new technologies unless the perceived benefits outweigh the perceived risks, and agencies may have concerns about validation and how the courts might interpret the statistics. Any new technology also requires training both the examiners as well as the finder of fact.

Because the pattern comparison disciplines lack a current procedure to calculate a likelihood ratio measure, there is no direct estimate of the overall strength of the evidence, which introduces the risk of examiners overstating or understating the strength of the forensic evidence when they use verbal statements to communicate the strength of the evidence. Most pattern comparison disciplines now have error rate studies, which estimate the likelihood that an examiner would make various errors (e.g. [30], but not the strength of the evidence for each image pair. Because of this, the language used to communicate the conclusion may not be calibrated relative to this strength of evidence. For example, novices typically view the term Identification to mean to the exclusion of all others [27] despite the fact that examiners have been taught to testify that Identifications cannot mean exclusion to all others [10]. Some error rate studies have collected difficulty measurements to acknowledge the differences that might exist across comparisons [7,32,33]. However, the complexity or difficulty of a comparison is typically not communicated to the factfinder unless the examiner provides additional nuance or context during reporting or testimony. Thus, all conclusions within the same reporting category are treated equivalently by the consumer.

In sum, the current conclusion scale used by fingerprint examiners is inappropriate in that it transposes the conditional, has only three categories, is not transparent, and does not directly communicate the differences in strength of the evidence for each image pair within a reporting category. Deriving a likelihood ratio solves all four of these problems and would allow the factfinder to appropriately weigh the information received from pattern comparison testimony with the other facts of the case.

The goal of this article is to create a quantitative likelihood ratio measure for the pattern comparison disciplines, here applied to fingerprint comparisons. Rather than rely on image features such as minutiae locations, we adopt an approach based on the consensus of human expert decisions. We are not using the minutiae, computerized quality metrics, or examiner annotations of the images at all, only the distribution of responses by experts in error-rate studies. Our approach uses a mathematical model that allows us to convert human judgments into quantitative likelihood ratios that provide estimates of the underlying evidentiary strength of each image pair as expressed by the collective responses of human experts. Our approach has three key assumptions that we will expand upon:

- The distribution of human decisions for an image pair in an error rate study is a proxy for the strength of the evidence for that image pair.
- The distribution of decisions can be summarized using an ordered probit model for each image pair.
- By combining the output of the ordered probit model with the known ground truth (mated or nonmated) for each pair in a database, we can compute a likelihood ratio that estimates the evidentiary strength of each image pair in our dataset.

Below we expand on these assumptions and apply the model to two extant datasets.

1. Summarizing human expertise

Absent a measure of the degree of correspondence between two impressions provided by a computer-based approach, the judgments of human experts are perhaps the best measure of the strength of support for the same-source and different-source propositions. Consider the image pair shown in Fig. 1, which is an image pair used by Busey, Klutzke, Nuzzi, and Vanderkolk [6]. That study collected data from 66 examiners who each conducted 60 comparisons using one of three types of scales (traditional, expanded traditional, or a strength of support scale). The goal of the article was to measure how the distribution of conclusions would change if examiners were given different types of conclusion scales. The traditional scale used just Exclusion, Inconclusive, and Identification, while the expanded traditional scale added Support for Different Sources and Support for Common Source. The strength of support scale also had five responses, but changed Exclusion to Extremely Strong Support for Different Sources and changed Identification to Extremely Strong Support for Common Source. Examiners were first asked whether the latent impression was of value (sufficient evidence to perform a comparison) or not, and if they selected 'no value' we did not include those trials in the present analysis.

The response distributions for the image pair in Fig. 1 are shown on the right panels of Fig. 1. On the top-right panel, examiners in the traditional scale were split, with 28 examiners making an

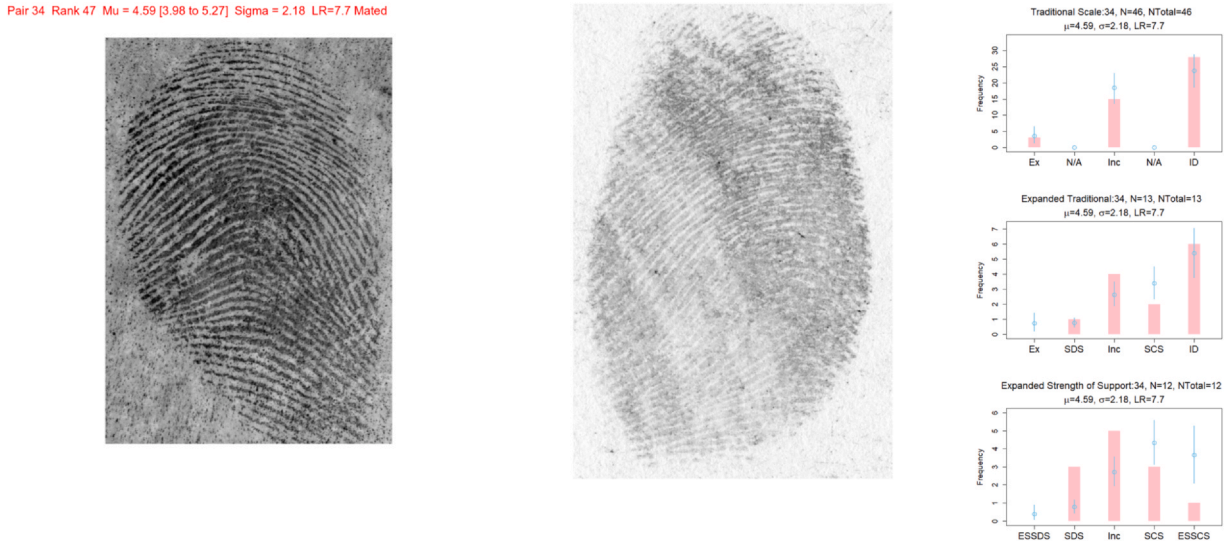

Fig. 1. Example comparison from Busey et al. [6]. This pair is mated. The distributions on the right correspond to the number of examiners who reached each conclusion. The top-right panel has data from the traditional scale (Ex is Exclusion, Inc is Inconclusive, and ID is Identification). The middle panel has data from an expanded traditional scale that included Support for Common Source (SCS) and Support for Different Sources (SDS); and the bottom panel contains data from a Strength of Support Scale that replaces Identification with Extremely Strong Support for Common Source (ESSCS) and Exclusion with Extremely Strong Support for Different Sources (ESSDS). The blue dots and lines are the model fits from the ordered probit model as described in the text.

Identification conclusion, 15 saying inconclusive, and three reaching an Exclusion conclusion. In the middle panel, the distribution of responses on the expanded traditional scale favor the Identification end of the scale with a mixture of Inconclusive, Support for Common Source, and Identification conclusions. In the lower panel, examiners were distributed across multiple conclusions. In all three scales there is variability across examiners, and it is this distribution of responses that we will use to characterize the strength of support for all images in our dataset.

Contrast the image pair in Fig. 1 against that in Fig. 2. The latent impression in Fig. 2 has much higher quantity and quality than the latent in Fig. 1, and the examiners in all three scales were unanimous in their conclusions in Fig. 2. The differences in response distributions between these two image pairs provide an intuitive sense that the observations that examiners derived from the second image pair provide more support for the same source proposition than the first image pair. Note, however, that 28 of 46 examiners reached an Identification conclusion in the traditional scale for the image pair in Fig. 1, and if two examiners in the same agency both reached the Identification conclusion then the image pair in Fig. 1 would be reported out as an Identification. Without additional explication, this conclusion would be treated as equivalent to the conclusion provided by the image pair in Fig. 2 by the factfinder. However, we argue that these two image pairs should provide different strengths of support for the same and different source propositions based on the response distributions provided by each image pair. A quantitative likelihood ratio better communicates these different strengths of support than a categorical conclusion response scale.

Complete response distributions for all three scales for the 60 fingerprint pairs are found in Table 1, which provide a sense of how the distribution of responses can vary across different image pairs. These are sorted by the likelihood ratio, which is described below, but generally the image pairs at the top of the table tended to be nonmated and garner mostly Exclusion or Extremely Strong Support

Fig. 2. Example comparison from Busey et al. [6]. This pair is mated, and the distributions on the right correspond to the number of examiners who reached each conclusion. See Fig. 1 for description of graphs.

Table 1
Response distribution from Busey et al. [6] sorted by likelihood ratio. The mu and sigma are from the ordered probit model. Subsequent columns represent counts of each response for each of the three scales. NV = Not of Value. Note that odd pairID values are nonmated and even pairID values are mated. Bold likelihood ratio values are those pairs in which examiners gave more Identification decisions than all other responses combined (including NV), which is one measure of whether a comparison might considered casework-like quality.

				Traditional				Expanded Traditional						Strength of Support					
pairID	mu	sigma	LR	Ex	Inc	ID	NV	EX	SDS	Inc	SCS	ID	NV	ESSDS	SDS	Inc	SCS	ESSCS	NV
43	-4.45	3.17	0.01	40	0	1	0	12	0	0	0	0	0	17	0	1	0	0	0
35	-4.09	1.12	0.01	37	0	0	0	18	0	0	0	0	0	16	0	0	0	0	0
51	-2.64	1.65	0.02	29	0	0	0	19	1	0	0	0	0	22	0	0	0	0	0
29	-2.52	1.63	0.02	36	0	0	0	11	1	0	0	0	0	20	0	0	0	0	0
57	-2.17	1.94	0.02	34	1	0	0	18	0	0	0	0	0	15	1	1	0	0	0
55	-1.88	1.29	0.03	38	0	0	0	17	0	0	0	0	0	16	1	0	0	0	0
11	-1.48	1.81	0.03	29	2	0	1	18	0	0	0	0	0	14	2	1	0	0	0
17	0.62	2.15	0.12	20	12	1	5	8	2	2	1	0	1	9	4	1	2	0	2
59	0.78	1.63	0.14	22	6	1	0	10	5	2	0	1	0	5	13	1	1	0	0
53	0.79	1.59	0.14	18	9	0	7	7	2	1	1	0	6	6	7	2	1	0	3
5	0.85	1.46	0.15	24	6	1	0	13	2	3	0	0	0	3	8	8	0	0	0
3	0.87	1.09	0.15	18	7	0	5	8	6	2	0	0	3	7	6	3	0	0	2
7	1.20	0.96	0.22	4	3	0	29	2	3	0	0	0	14	1	2	1	0	0	9
37	1.22	0.82	0.22	25	8	0	4	1	7	3	0	0	2	3	7	4	0	0	6
25	1.24	1.42	0.23	20	14	0	0	5	2	10	1	0	0	8	6	3	0	0	0
1	1.26	1.61	0.23	18	8	0	0	9	4	5	2	1	0	5	9	8	1	0	0
13	1.32	0.65	0.25	5	2	0	24	3	3	0	0	0	11	0	6	3	0	0	14
31	1.33	1.39	0.25	14	16	0	1	5	5	8	0	0	1	11	2	7	0	0	1
9	1.36	1.15	0.26	18	21	0	0	5	2	4	0	0	1	6	8	4	0	0	0
41	1.40	1.31	0.28	18	16	0	0	7	5	7	0	0	0	6	4	7	1	0	0
19	1.51	0.75	0.32	12	14	0	17	1	4	1	0	0	8	1	5	2	0	0	4
23	1.61	0.69	0.36	8	22	0	11	6	5	0	0	0	5	0	6	3	0	0	4
49	1.68	1.35	0.39	16	14	0	0	4	4	9	0	0	1	4	5	11	0	1	0
45	1.82	1.01	0.46	9	20	0	7	6	2	3	0	0	6	1	3	10	0	0	5
47	1.87	0.95	0.49	7	23	0	1	3	9	5	0	0	0	4	8	12	1	0	0
21	1.90	0.94	0.51	13	26	0	0	3	5	8	0	0	2	2	2	9	0	0	0
38	2.30	0.68	0.79	2	20	0	15	1	1	9	0	0	9	0	3	7	0	0	7
2	2.38	1.05	0.86	7	20	0	13	1	3	6	1	0	5	0	0	5	2	0	9
15	2.40	0.69	0.87	2	27	0	8	0	0	13	0	0	2	1	4	11	0	0	3
39	2.42	0.93	0.89	1	4	0	30	0	0	1	0	0	15	0	0	2	0	0	17
22	2.46	0.63	0.92	0	16	0	11	2	1	11	0	0	5	0	1	14	0	0	8
10	2.46	0.67	0.92	2	32	0	1	1	2	9	0	0	1	0	3	19	1	0	1
33	2.66	0.89	1.11	2	27	0	9	0	4	8	3	0	2	0	3	6	2	0	5
56	2.70	0.99	1.14	4	24	0	7	0	4	7	4	0	3	0	3	10	4	0	1
50	2.85	1.10	1.30	2	11	1	18	1	1	4	2	0	10	0	0	8	2	0	12
12	2.93	1.41	1.39	3	20	5	1	1	3	7	5	2	1	1	5	11	2	0	0
27	3.02	1.50	1.50	0	1	0	38	0	0	0	0	0	14	0	0	0	0	0	14
24	3.24	1.11	1.83	1	10	2	15	1	0	8	4	0	7	0	1	7	6	0	8
44	3.43	1.47	2.19	2	20	5	6	1	1	4	8	1	2	0	4	6	6	2	2
6	3.54	0.82	2.43	0	35	3	4	0	0	5	7	0	1	0	0	8	3	1	3
26	3.54	1.32	2.44	2	21	8	3	0	1	7	3	1	2	0	2	8	9	1	0
48	3.64	1.14	2.71	0	26	4	4	1	2	5	11	1	1	0	0	7	7	2	2
40	3.99	1.66	3.94	3	20	13	0	1	2	3	7	3	0	0	2	1	13	2	0
46	4.02	1.14	4.06	2	20	9	1	0	1	2	12	3	1	0	0	4	14	0	0
60	4.34	1.87	5.83	2	15	14	3	1	0	3	5	6	3	0	2	2	7	2	4
14	4.38	4.01	**6.10**	7	7	16	0	5	0	3	2	12	0	3	4	2	4	4	0
34	4.59	2.18	**7.71**	3	15	28	0	0	1	4	2	6	0	0	3	5	3	1	0
32	5.37	2.28	18	1	8	16	10	0	1	0	1	5	5	0	1	3	5	6	5
28	6.06	1.98	**44**	0	6	23	6	0	1	2	1	11	1	0	0	1	7	9	2
52	6.47	2.28	**78**	1	6	22	3	0	1	0	4	11	2	0	0	0	5	13	0
30	6.93	1.68	**160**	0	2	30	1	0	0	1	1	16	0	0	0	0	5	13	0
16	7.24	2.57	**267**	0	6	30	4	0	0	0	0	10	0	1	0	0	5	12	0
54	7.73	2.80	**643**	0	7	33	1	0	1	0	0	13	4	0	0	1	0	9	0
4	8.23	3.05	**1649**	2	3	34	0	0	0	0	1	17	1	0	0	0	3	9	0
36	8.45	3.13	**2543**	0	3	32	0	0	1	0	1	12	0	1	0	0	3	17	0
58	8.94	4.76	**7110**	2	2	36	0	1	2	0	1	12	0	2	0	1	1	11	0
18	9.38	2.33	**17628**	0	0	35	0	0	0	0	1	16	0	0	0	1	0	19	0
8	9.75	3.34	**38689**	0	0	21	0	1	1	0	1	22	0	0	0	0	2	21	0
42	10.24	4.75	**103527**	4	0	31	1	0	0	1	0	14	0	0	1	0	1	15	0
20	10.30	1.13	**117647**	0	0	37	0	0	0	0	0	16	0	0	0	0	0	17	0

for Different Sources responses. Looking down the table, the pairs start to receive more Inconclusive responses and then finally more Identification or Extremely Strong Support for Common Source responses at the very bottom. Thus, we see a continuous shift in support for the different source proposition at the top of the table all the way to support for the same source proposition at the bottom. The next step is to summarize across the three scales to provide a single estimate of the strength of support provided by each image pair, which will explain the mu, sigma, and LR columns in Table 1.

1.1. Ordered probit model

To summarize the distributions of responses across examiners (and across scales), we used an ordered probit model. This model gets its name from the fact that we consider the responses from

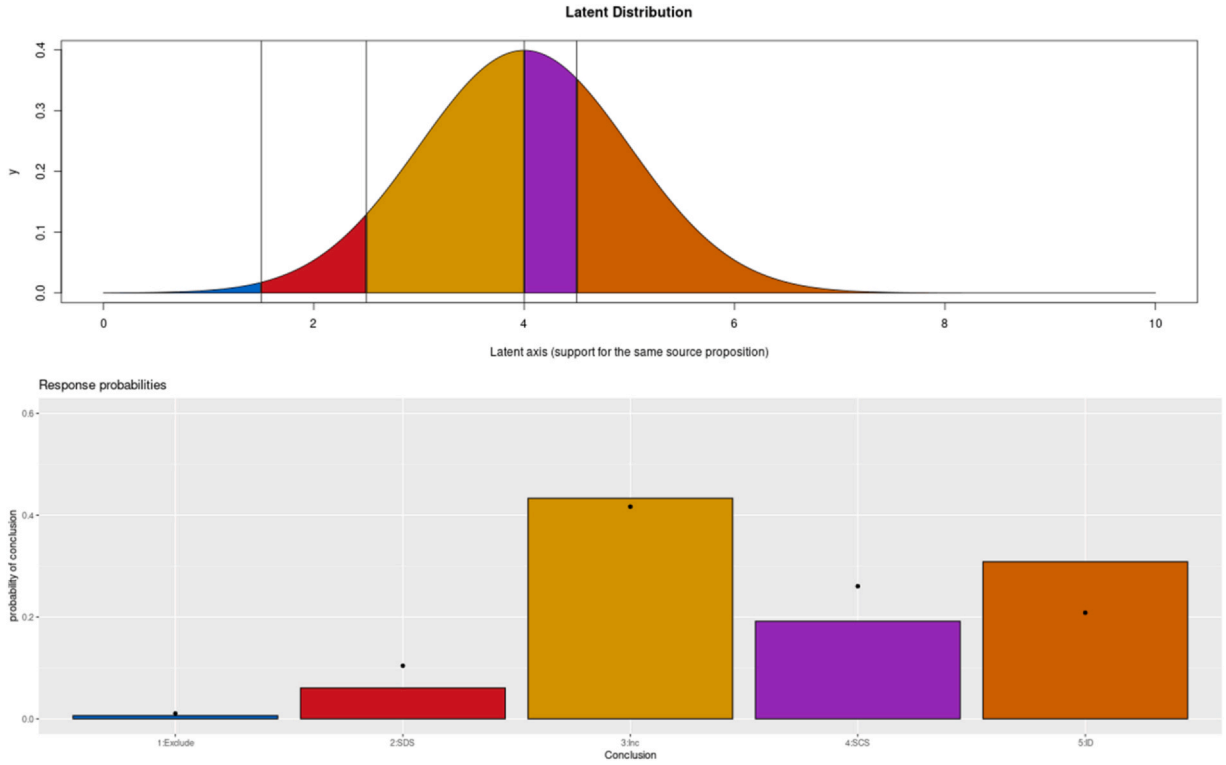

Fig. 3. Illustration of the ordered probit model applied to hypothetical fingerprint response probabilities (black dots in lower panel) with a poor model fit. Upper panel: the ordered probit model assumes a normal distribution that reflects the collection of internal responses from all of the examiners who completed a comparison on a given image pair. Four decision thresholds are placed along the latent axis to produce predictions for the frequency for each of the five conclusions in this expanded scale. The area under the normal distribution between different thresholds determine the predicted response frequencies for each conclusion. Lower panel: the predicted response frequencies for each conclusion is given by the height of the colored bar, and are computed directly from the corresponding areas in the top panel. SDS corresponds to Support for Different Sources, and SCS corresponds to Support for Common Source.

examiners on an ordinal scale, in the sense that Identification provides more support for the same source proposition than Inconclusive does, and Inconclusive provides more support for the same source proposition than Exclusion does. The term 'probit' comes from the assumption of the normal distribution along a latent (hidden) axis that is meant to summarize all of the responses to that image pair. Note that the use of 'latent' in the statistical context means that it cannot be directly observed but can be inferred using a modeling approach and should be distinguished from the shorthand term 'latent' in friction ridge comparisons where it refers to a friction ridge impression typically obtained from a crime scene. To disambiguate these, we will refer to the statistical concept of latent by referencing it as a dimension, and sometimes remind the reader that this is hidden. The concept of a latent impression will always include the term impression.

Fig. 3 illustrates the elements of the ordinal probit model, and below we describe the foundational assumptions and then give an intuitive explanation for the model. The reader is invited to try two interactive demonstrations of the model that we created for the purposes of this paper. The first models the traditional scale:

https://iupbsapps.shinyapps.io/OrderedProbitDemoTraditional/.

while the second is a representation of an expanded traditional scale:

https://iupbsapps.shinyapps.io/OrderedProbitDemo/.

In each demo, adjust the sliders to make the bars correspond to the black dots (which represent a hypothetical examiner response distribution). See the text in the apps for more instructions and tips.

The ordered probit model assumes that there exists an underlying latent (hidden) dimension along which the examiner accumulates information or evidence (top panel of Fig. 3). At the conclusion of a comparison, each examiner will end up at some final value along this latent axis. The endpoints of the latent dimension might be something like: *The most support imaginable for the different source proposition* and *the most support imaginable for the same source proposition*, although in principle the axes are unbounded. We cannot typically measure this final value using the categorical system, but for the sake of exposition, imagine that we had some brain recording device that could measure a value on a continuous scale that reflected the accumulated evidence for the two propositions along the latent axis. Of course, the examiner has access to this final latent value because they make their decision based on this final value.

The second assumption of the ordered probit model is that there is a series of decision thresholds or criteria that are placed along the latent (hidden) dimension to obtain one of the pre-approved conclusions (i.e. Exclusion, Inconclusive, or Identification). For example, if the final value along the latent dimension exceeds the threshold separating Inconclusive from Identification, an examiner would report an Identification conclusion. These thresholds can be thought of as system-wide or consensus thresholds that might not exactly correspond to the thresholds for a given examiner, but instead reflect the output of the group. Each conclusion scale has a different set of thresholds as described in more detail below.

The third assumption is that the distribution of final latent values across examiners can be summarized with a normal distribution that has a particular mean and standard deviation. This is shown in the top panels of Fig. 3 and Fig. 4, with the thresholds defining different colored regions. The assumption of normality is widely supported [14], although other distributions are possible. We return to this topic in the Discussion section where we test other latent distributions.

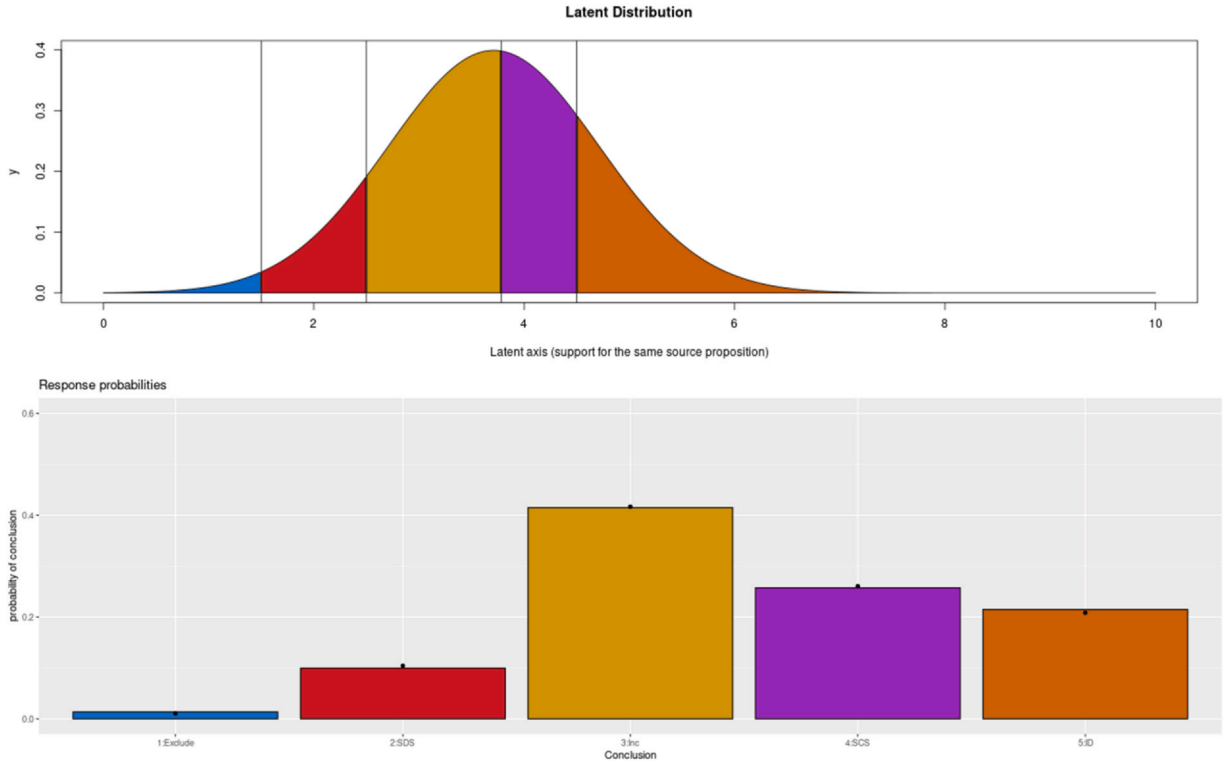

Fig. 4. Illustration of an ordered probit model applied to hypothetical fingerprint response probabilities (black dots in lower panel) with the best fitting parameters. Compare with Fig. 3 to see how shifting the normal distribution to the left slightly and shifting the upper threshold to the right allows the predicted response probabilities to align much closer to the hypothetical response probabilities.

The fourth assumption is that the predicted proportion of examiner responses for a given conclusion is given by the area under the normal distribution between the associated thresholds. This relation can be observed by adjusting the sliders in the app and watching how the different colored areas change, or by comparing Fig. 3 to Fig. 4. The area of each colored region directly determines the height of the columns in the lower graphs of Fig. 3 and Fig. 4, and the heights of the bars can be compared against the hypothetical distribution of examiner responses (black dots) to find the parameters that best account for the examiner responses.

More formally, and using the notation from equation 23.1 in chapter 23 of Kruschke [11], the probability of outcome k (Exclusion, Inconclusive, or Identification in the case of the traditional scale) is:

$$p(y = k|\mu, \sigma, \{\theta_j\}) = \Phi((\theta_k - \mu)/\sigma) - \Phi((\theta_{k-1} - \mu)/\sigma) \quad (1)$$

where $\Phi(z)$ represents the cumulative area under the standardized normal, $\theta_k$is the k^{th} threshold, and μ and σ are the mean and standard deviation of the normal distribution on the latent dimension. The lowest and highest regions are special cases where we assume virtual thresholds at $-\infty$ and $+\infty$ (see the above chapter for more details).

The goal of the ordered probit model is to summarize the support for the same source and different source propositions based on the distribution of examiner responses to an image pair. The mean μ of the normal distribution represents the support for the two propositions provided by the examiner responses, with larger values of μ corresponding to more support for the same source proposition. The standard deviation σ represents the consistency across examiners. While the decisions that examiners make are not numeric, the underlying latent dimension is on a metric scale, which is an assumption of the ordered probit model; we are using the model to bootstrap our way from an ordinal response scale to a metric underlying dimension that will characterize the strength of evidence for the two propositions. Experimentation with the two apps is intended to demonstrate that changes in any one parameter (slider) will produce changes in at least two and often all of the predicted response probabilities, and that the relation between these changes and the response probabilities is often unintuitive. The goal is to adjust the parameters (sliders in the apps) to make the predicted response frequencies correspond to the observed response frequencies. To automate this process for all of our data, we turn to computer-based approaches using a Bayesian framework.

1.2. Fitting the ordered probit model to extant data

The goal of our modeling is to summarize the responses from all of the examiners who compared a particular pair of impressions, and do so regardless of which scale they used. We use Bayesian methods described in Chapter 23 of Doing Bayesian Data Analysis [11]. We fit the response frequency data in Table 1, using only responses in which the examiner did not rate the latent as No Value. We fit the ordered probit model to our ordinal response frequency data with the following assumptions:

- To establish the scale of the underlying latent distribution, the thresholds for the traditional scale trials were set to 1.5 and 4.5 on the latent scale. We chose 1.5 and 4.5 because our expanded scales have 5 categories, and this approach easily generalizes to scales of different sizes. Note that this is not a critical assumption, and we could have chosen other values with identical results.
- Prior research demonstrated that examiners will shift their thresholds when additional categories are added, such as in the expanded traditional or strength of support scales [6,8]. Therefore, we allowed the four thresholds for each of the two expanded scales to freely vary. The thresholds are still interpretable relative to the fixed thresholds from the traditional scale, which sets the scale of the latent axis along which the expanded scale thresholds are interpreted.

- The thresholds for the two expanded scales are estimated across all 60 pairs in the dataset.
- Each of the 60 pairs has an individually-estimated mean μ and standard deviation σ for the normal distribution on the latent dimension. The mean and standard deviation are jointly estimated across all three scales (observe that the mean μ and standard deviation σ in the three graphs in Fig. 1 are identical).
- The standard deviations are subject to shrinkage (see Kruschke [11], which essentially uses other standard deviations as the prior for each standard deviation, and tends to reduce the variation across the standard deviations. This addresses concerns where examiners were unanimous for an image pair, which does not provide strong constraints on either μ or σ.
- We used noncommittal (diffuse) prior distributions for the means and standard deviations for each of the n pairs, and used a gamma distribution for the higher-level distribution of the standard deviations. For the i^{th} pair:

$$\mu_i \sim normal(mean = (1 + 5)/2, sd = 1/(5^2))$$

$$\sigma_i \sim gamma(\sigma_{Shape}, \sigma_{Rate})$$

$$\sigma_{Shape} = 1 + \sigma_{Mode} * \sigma_{Rate}$$

$$\sigma_{Rate} = (\sigma_{Mode} + \sqrt{\sigma_{Mode}^2 + 4 * \sigma_{SD}^2})/(2 * \sigma_{SD}^2)$$

$$\sigma_{Mode} \sim gamma(mode = 3.0, sd = 3.0)$$

$$\sigma_{SD} \sim gamma(mode = 3.0, sd = 3.0)$$

- Each estimated k^{th} threshold θ_k from the expanded scales had a normal prior with noncommittal values:

$$\theta_k \sim normal(mean = k + .5, sd = 1/(2^2))$$

- We used Monte Carlo Markov Chain estimation to find parameter sets that produce a distribution of posteriors with highest credibility, and used the median of this distribution for the estimate of the most credible mean and standard deviation.

The parameters of the model were estimated using MCMC methods using the JAGS package [21,22] in R (Team, 2013). We used 500 initialization steps and 1000 adaptation steps, and a final chain of 60,000 steps thinned every 5 steps. All effective sample sizes were above 10,000 and the chains showed little evidence of autocorrelation. Further details and all analysis code are found in the osf.io site: https://osf.io/tmgdn/?view_only= 028e61dccd984d0a95f8107b8543aee3.

To evaluate the adequacy of the fit of the ordered probit model, the fits to all response distributions are shown in Fig. 5 for the traditional scale, Fig. 6 for the expanded traditional scale, and Fig. 7 for the strength of support scale. The results of this analysis are predictions for the response frequencies for each pair of impressions in our dataset. These predictions show as the blue dots in Fig. 5, Fig. 6, and Fig. 7, and the blue lines denote the 95% highest density interval (HDI) on the frequency estimates. The model is doing a remarkable job given that across the three graphs there are 10 degrees of freedom for each image pair, which are being accounted for by two free parameters (the mean μ and standard deviation σ). The four thresholds from the expanded traditional scale and the four thresholds for the strength of support scale are also free parameters but they are fit across all 60 pairs and are therefore subject to a large number of constraints. Thus, the model is very far from being saturated, and some of the mis-predictions of the model are coming from simple multinomial variance.

The goal of the ordered probit model is to characterize the rich response data obtained in the simulated casework using only a few parameters. The mean μ reflects the location of the normal distribution along the latent dimension. The μ in Fig. 1 is 4.59, while the μ in Fig. 2 is 10.27. This reflects the fact that more examiners felt comfortable using the highest category of responses for the image pair in Fig. 2 and therefore provides more support for the same source proposition and less support for the different sources proposition than the image pair in Fig. 1. The standard deviation σ is a measure of examiner consistency, and σ in Fig. 1 is 2.18 while σ in Fig. 2 is 1.12, meaning that examiners were more consistent in their responses for the image pair in Fig. 2.

1.3. Computing likelihood ratios

In this next section we describe how our approach allows us to compute likelihood ratios for individual image pairs based on the distribution of responses as summarized by the ordered probit model, combined with information from the rest of the image pairs and ground truth. The normal distribution for each image pair is summarized by its parameters mean μ and standard deviation σ. The parameters for each of the 60 image pairs define different normal distributions, and these are illustrated in Fig. 8, with a linear y axis on the left panel and a log y axis on the right panel. Each light red curve corresponds to the latent distribution of a nonmated pair, and each light blue curve corresponds to the latent distribution of a mated pair. The light blue and light red curves at different locations along the latent axis reflect the fact that different image pairs provide different amounts of support for the same source proposition (see Fig. 1 and Fig. 2 for examples of pairs that differ in their support for the same source proposition).

The height of the normal distribution at each location along the latent axis can be thought of as the relative likelihood that an examiner would have obtained a particular value along the latent axis for that image pair. Stated another way, the height of a given curve is the relative likelihood of observing a particular latent value given that image pair was selected. This is the definition of a probability density function, which is what these curves represent.

For a particular value along the latent axis, what is the relative likelihood that *any* mated image pair would result in that value? If we assume that the decisions for each image pair are independent, we can use the mutual exclusivity rule in probability to add together the mated distributions (light blue curves) to create the thick blue curve in Fig. 8, and add together the nonmated distributions (light red curves) to create the thick red curve in Fig. 8. These thick blue and red curves are each normalized so that each area sums to 1.0 to make them true probability densities. The height of the thick blue curve in Fig. 8 represents the relative likelihood of observing a particular latent value in the dataset *given* the same source proposition (i.e. mated), while the height of the thick red curve represents the relative likelihood of observing a particular latent value *given* the different sources proposition (i.e. nonmated).

The likelihood ratio is simply the ratio of these two relative likelihood curves (the thick blue curve divided by the thick red curve at every location along the latent axis). We plot the likelihood ratio as a function of the latent axis in Fig. 9, with the highlighted region corresponding to those image pairs that received majority ID decisions in the three conclusion scale. The relation between Fig. 9 and Fig. 8 can be seen in the right panel of Fig. 8, because in the right panel of Fig. 8 the distance between the red and blue curves is the log of the likelihood ratio. Thus, even though the thick red and thick blue curves appears to converge on the right side in the left panel of Fig. 8, the right panel of Fig. 8 illustrates that the two thick curves continue to diverge for larger values of the latent axis, which

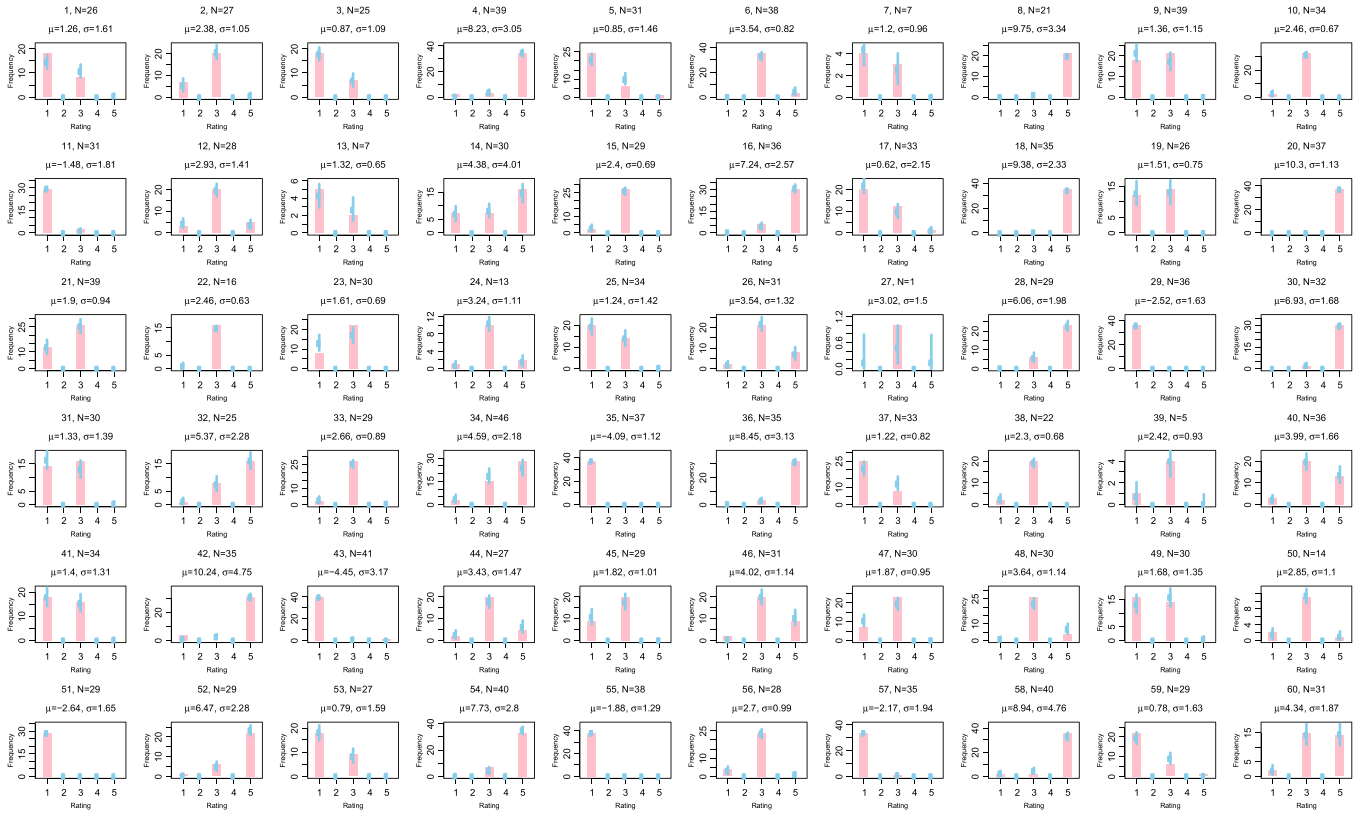

Fig. 5. Response frequencies (pink bars) and ordinal probit model fits (blue dots) for 60 comparisons from examiners who used the traditional scale from the Busey et al. [6] data. Note on these graphs a rating of 1 corresponds to Exclusion, 3 corresponds to Inconclusive, and 5 corresponds to Identification, although the model does not treat these as metric, only ordinal.

produces the larger likelihood ratios for larger values of the latent axis in Fig. 9.

A likelihood ratio of 1 implies equal support for the two propositions. On our latent axis, the thresholds of the traditional scale were set to 1.5 and 4.5, implying that a value of 3 should correspond to the midpoint of the latent scale, or equal support for the two propositions. The derived value of the likelihood ratio at a value of 3 along the latent axis is indeed very close to 1.0, suggesting that examiners treat the value midway between the two thresholds on the traditional scale as approximately equal support for the two propositions.

1.4. Computing likelihood ratios for individual image pairs

Having defined the likelihood ratio as a function of the values on the latent axis in Fig. 9, we can now use that relation to compute likelihood ratios for individual image pairs. We assume that the estimated value of μ for a given image pair is representative of the typical strength of support provided by the collection of examiners who completed a comparison on that image pair. The likelihood ratio for a pair is simply the height of the thick blue curve from Fig. 8 at its value of μ, divided by the height of the thick red curve from Fig. 8 at the value of μ. Because the two curves are computed at discrete intervals, we use linear interpolation of the thick red and thick blue curves to compute the likelihood ratio. Simply put, to find the likelihood ratio for any image pair, use the μ associated with that image pair to read off the height of the curve in Fig. 9. Note that the value μ was chosen to represent the typical evidentiary strength of a particular image pair. In most forensic testimony, only a single likelihood ratio is reported rather than a range. The MCMC process calculates a 95% highest density interval (HDI) for each image pair for the μ parameter, and this could be used to produce a range of likelihood ratios rather than a single value. However, a range of values may suggest a form of uncertainty to the fact finder that might obscure the correct interpretation of the values.

Table 1 provides the distribution of responses, μ, σ, and likelihood ratio values for all pairs in the dataset, sorted by the likelihood ratio values. Inspection of Table 1 gives a sense of how a particular distribution of examiner responses is associated with different likelihood ratio values. In the Discussion we provide some interpretation of these values.

Although we can consider the likelihood ratio values for nonmated and mated pairs in Table 1, operationally the only values that typically matter are for image pairs that are of casework-like quality and therefore would get reported as a conclusion to the court system. If half or more of examiners determined an image pair was an “Identification”, the image pair might reasonably be reported by an agency as an “Identification”. We defined these image pairs as “casework-like quality” and highlight these rows in Table 1. These image pairs demonstrate likelihood ratios ranging from about 10–100,000, with μ values between about 4.4 and 10 along the latent axis. The Supplementary Materials contained at the osf.io site linked above have a folder called *ImagesCombinedAndSorted*, which have all 60 image pairs along with the response distributions and likelihood ratios. These images will provide an important link between the computed likelihood ratios and the subjective estimate of comparison complexity. The reader is encouraged to view the images in this folder to get a sense of the likelihood ratios associated with image pairs of different perceived complexities. The blue region in Fig. 9 illustrates the approximate range of values for casework-like impressions.

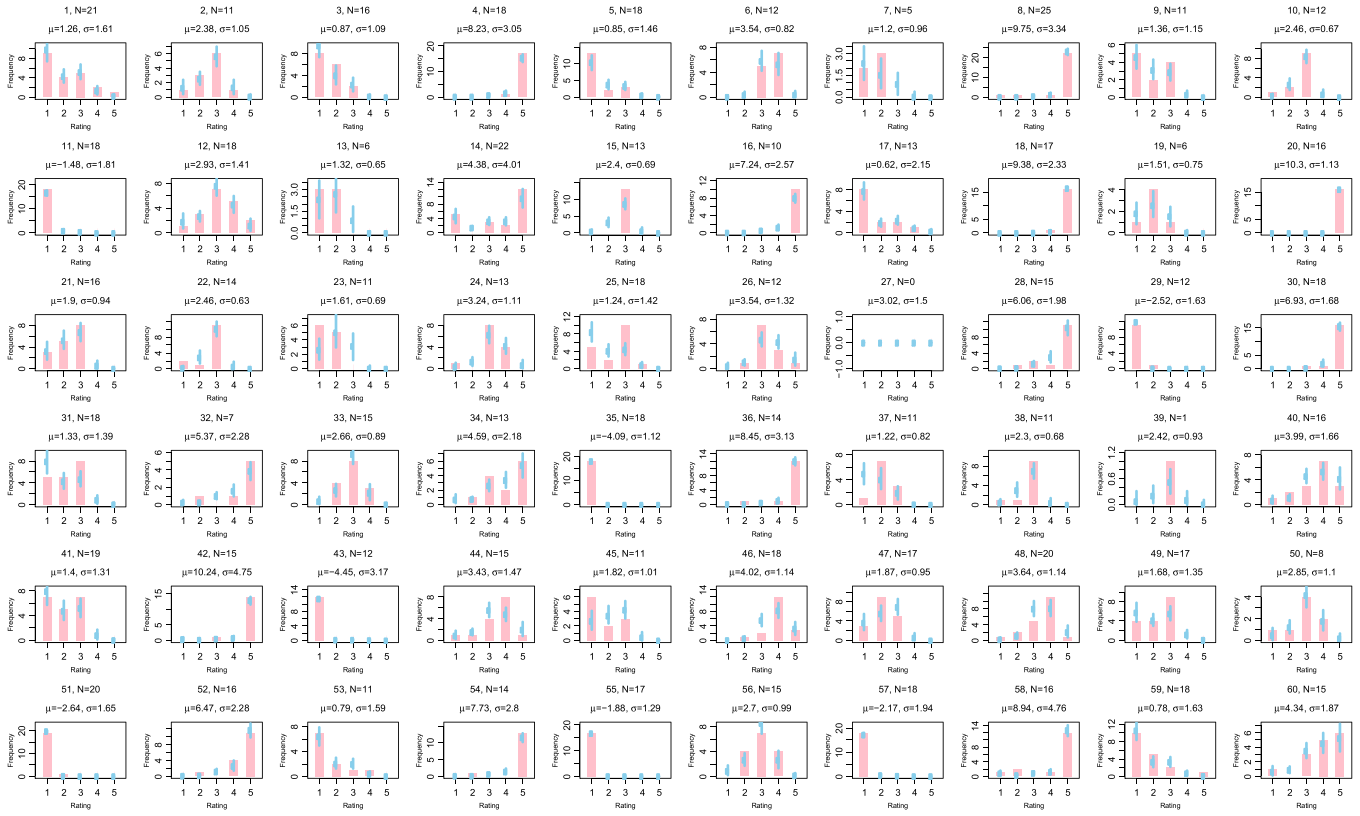

Fig. 6. Response frequencies (pink bars) and ordinal probit model fits (blue dots) for 60 comparisons from examiners who used the expanded traditional scale from the Busey et al. [6] data. Note on these graphs a rating of 1 corresponds to Exclusion, 2 corresponds to Strong Support for Different Sources, 3 corresponds to Inconclusive, 4 corresponds to Strong Support for Common Source, and 5 corresponds to Identification, although the model does not treat these as metric, only ordinal. Cells with no data (i.e. all examiners reported no value for that comparison in this scale) are still constrained by data in the other scales for that comparison.

1.5. Fitting the FBI/Noblis Black Box data

The above analyses were fit to the data obtained in the Busey et al. [6] research study, and while the use of expanded scales provides for better estimation of the parameters of the ordered probit model and the image pairs are available for publication, that study was not a true error rate study. To compute likelihood ratios from casework-like comparisons, we fit the data provided by the FBI/Noblis Black Box study [30] using the same ordered probit model as used above.[1] The Black Box study is seen as the gold standard for fingerprint studies designed to mimic casework [20] and 85% of the participants felt that the overall difficulty of the image pairs was similar to casework. The methods were designed to follow the typical procedures of casework and the data is therefore viewed as a reasonable proxy for operational performance. The study tested 169 latent print examiners who completed a total of 17,121 comparisons across 744 pairs of ground-truth images using a traditional 3-conclusion scale. The study replaced the word "Identification" with "Individualization" to reflect a brief trend at the time, although it is likely that examiners treated the two synonymously. As with the previous traditional scale dataset, we fixed the threshold separating Exclusion from Inconclusive at 1.5 and the threshold separating Inconclusive from Individualization at 4.5. The choice of values for these fixed thresholds is arbitrary and will not affect the likelihood ratios, and using the same values as in the prior model fits does allow for comparisons of μ and σ with the previous data set.

Examiners in this study could also rate each latent as "not of value" and therefore not be required to produce a conclusion. A substantial number of pairs in this study have fewer than 16 examiners who were willing to reach a conclusion. We judged the data from these image pairs to be too limited to include in the analyses, and therefore pruned these pairs. This left us with 491 pairs for our analyses. As with the previous dataset, each image pair has a mean μ that corresponds to the amount of support for the same source proposition, and a standard deviation σ that corresponds to the amount of agreement among examiners who completed that comparison. The Markov Chain Monte Carlo model demonstrated clear convergence with virtually all effective sample size values above 10,000 and all above 6000. There was little evidence for autocorrelation in the chains.

Table 2 provides the response distributions for the FBI/Noblis Black Box study for every 10th pair and the full table is found on the osf.io site. As with Table 1, the rows are sorted by the likelihood ratio, and pairs toward the top of the table tend to be nonmated and garnered mostly Exclusion responses. Pairs lower down have more Inconclusive responses, and pairs toward the bottom have more Individualization responses. The lower rows tend to have larger μ values and larger likelihood ratios (discussed below).

Fig. 10 presents the response distributions for 60 randomly-selected image pairs from the Black Box study. The names indicate the ground truth (M for mated and N for nonmated). As might be anticipated by the fact that each graph has two degrees of freedom and two free parameters, the model predictions in blue are quite accurate. This demonstrates that the model is accurately capturing the response distributions and is a reasonable proxy for the consensus strength of evidence as measured by the collective response behavior of the examiners. However, the model is fully saturated because

[1] The data from the FBI/Black Box study can be downloaded from https://www.fbi.gov/services/laboratory/scientific-analysis/research-and-support/black-box-study-results

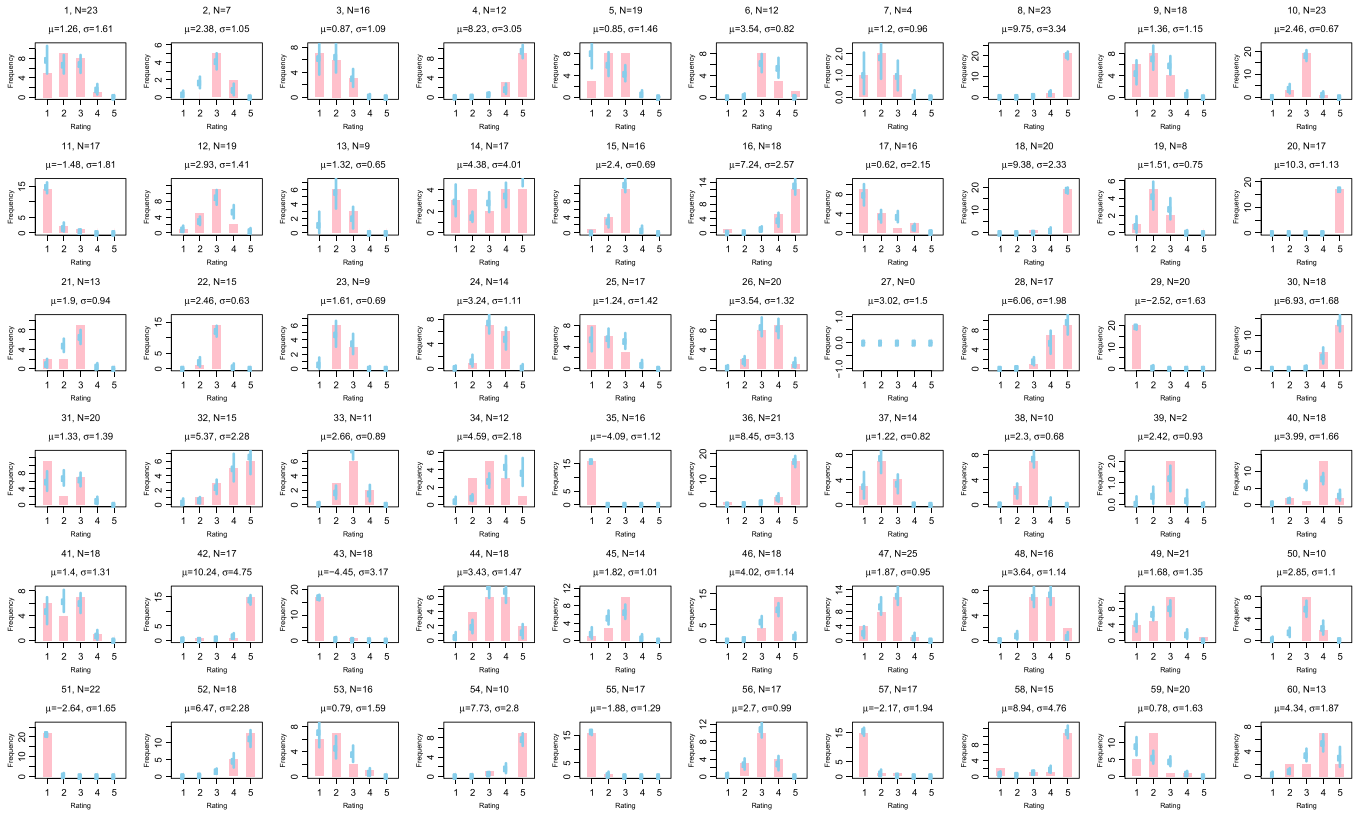

Fig. 7. Response frequencies (pink bars) and ordinal probit model fits (blue dots) for 60 comparisons from examiners who used the strength of support scale from the Busey et al. (2022) [6] data. Note on these graphs a rating of 1 corresponds to Extremely Strong Support for Common Source, 2 corresponds to Strong Support for Different Sources, 3 corresponds to Inconclusive, 4 corresponds to Strong Support for Common Source, and 5 corresponds to Extremely Strong Support for Common Source, although the model does not treat these as metric, only ordinal. Cells with no data (i.e. all examiners reported no value for that comparison in this scale) are still constrained by data in the other scales for that comparison.

there are an equal number of free parameters and degrees of freedom, and so these good fits are to be expected.

Fig. 11 shows the relative likelihood of observing a given latent value for each mated (light blue curves) or nonmated (light red curves) image pair. As with Fig. 8, the thick red curve corresponds to the normalized sum of the light red curves (non mated trials), and the thick blue curve corresponds to the sum of the light blue curves (mated trials). The shape of these curves is quite similar to those in Fig. 8. The clusters of light red and light blue curves at the extremes correspond to image pairs where the examiners were unanimous and there are more of these image pairs than in the Busey et al. [6] study, in part due to the use of the three-conclusion scale where fewer categories makes it easier to create unanimity among examiners.

The ratio of thick blue to thick red curve values at each point along the latent axis is the likelihood ratio, which is plotted in

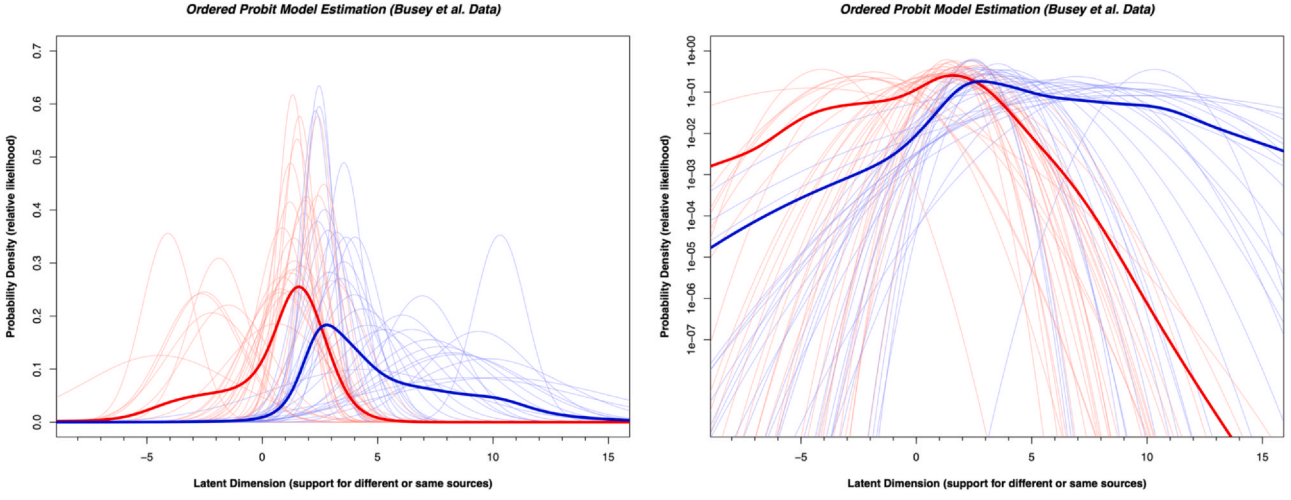

Fig. 8. Left panel: Relative likelihood of observing a given latent value for each mated (light blue curves) or nonmated (light red curves) image pair from the Busey et al. [6] data. The parameters for each normal distribution were derived from the ordered probit model fit to all three scales for each comparison. The thick red curve corresponds to the sum of light red curves, normalized to have an area of 1.0. It represents the relative likelihood of observing each value of the latent axis from *any* nonmated comparison. The thick blue curve represents the relative likelihood of observing each value of the latent axis from *any* mated comparison. Right panel: Same data plotted on a log (base 10) axis.

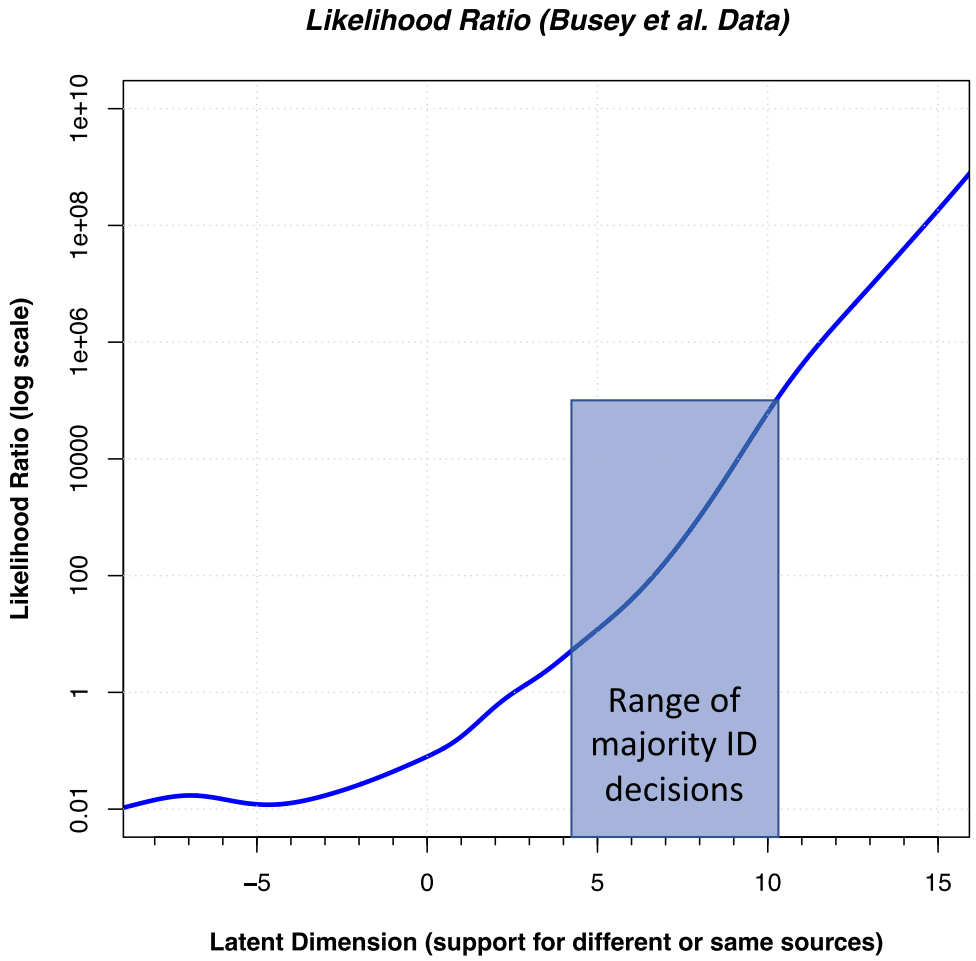

Fig. 9. Likelihood ratio values for different values along the latent axis from the Busey et al. [6] data. The y-axis is plotted on a log(10) axis. The log of the likelihood ratio can be observed directly as the difference between the thick blue and thick red curves in the right panel of Fig. 8. The blue area represents the range of image pairs that received a majority of identification decisions.

Fig. 12, with the estimated range of majority ID decision image pairs highlighted. As with Fig. 9, these values are plotted on a log ordinate axis. As with the previous dataset, we can compute a likelihood ratio for each image pair in the dataset by using the μ value for that pair to find the height of the red and blue curves at that μ value, the ratio of which gives the likelihood ratio for that image pair. These are shown in Table 2 for a subset of the images, and the full table is found in the supplementary information.

The bolded likelihood ratio values in Table 2 had more Identification responses than all other conclusions combined, which might be a rough proxy for casework-like quality. In the full table these likelihood ratios demonstrate a range of 49–22,414. These image pairs were associated with μvalues ranging from 4.9 to 10. Inspection of Fig. 12 gives a rough range of likelihood ratios of 50–20,000 for this range of μ values. This is a slightly narrower range than the values observed in the previous dataset, Fig. 9, which ranged from 10 to 100,000.

1.6. Parameter sensitivity analysis

We conducted a sensitivity analysis to determine whether the range of likelihood ratio values are a function of our particular choice of model parameters. There are three model assumptions that have the potential to affect the range of observed likelihood ratio values.

The first is the assumption of the normal distribution for the underlying latent distribution. If some examiners are outliers and behave differently than their colleagues, this might be better reflected by underlying distributions that have heavier tails. To model this, we replaced the normal distribution on the latent dimension with a Student's T distribution with five degrees of freedom, which produces a fairly heavy-tailed distribution. The net result was that the likelihood ratio values were reduced, with no likelihood ratios above 1100 for either dataset. This likely results from the fact that heavier tails in the distributions simply allow the nonmated pair distributions to extend further to the right, lowering the likelihood ratio values. We also replaced the normal distribution with a logit distribution and found similar results as above, which might be expected based on the fact that the logit also has a fairly heavy-tailed distribution.

The second potentially consequential assumption is the prior on the values of μ. Note that in Fig. 11 there are a group of distributions centered at -3 and 9. These come from the pairs that had unanimous responses, and the assumption of a fairly narrow prior on the expectation for the distribution of μ values will affect the location of these two groups along the latent axis. This results from the fact that when the data is unanimous, the model has no constraints and in principle could keep pushing the distributions to the extremes trying to get as much area under the normal distribution to the right of the upper threshold or to the left of the lower threshold. This is constrained in the Bayesian approach through the use of the prior on μ, which sets our expectations for how extreme the μ values should be. The previous models used a standard deviation of 5, which is the number of categories in the expanded scales. However, to explore the dependence of the likelihood ratios on this assumption, we fixed

Table 2
Representative data (decimated) from the FBI/Noblis Black Box study. The mu and sigma are from the ordered probit model. Pairs are sorted by the likelihood ratio (LR). Counts on the right side of the table correspond to the number of examiners who gave Exclusion (EX), Inconclusive (Inc), Individualization (ID), or No Value (NV) responses. Ground truth for each pair is given by the first letter of pairID (N = nonmated; M = mated). Bold likelihood ratio values are those pairs in which examiners gave more Identification decisions than all other responses combined (including NV), which is one measure of whether a comparison might considered casework-like quality. This is a decimated table of the full table which is found in the Supplementary Information as BlackBoxDataForTable2.csv. Note that tradeoffs between μ and σ can produce larger values of μ for pairs that have one or more erroneous exclusions than those that do not (see bottom rows), although this has only a small effect on the likelihood ratio values.

pairID	mu	sigma	LR	EX	Inc	ID	NV
N134386	-3.30	3.42	0.01	21	0	1	0
N141140	-3.06	0.91	0.01	24	0	0	0
N052071	-3.04	0.91	0.01	25	0	0	0
N316304	-3.03	0.91	0.01	25	0	0	0
N352366	-3.01	0.90	0.01	23	0	0	0
N030028	-3.00	0.91	0.01	23	0	0	0
N143447	-2.35	3.33	0.01	20	1	1	0
N232229	-1.37	1.65	0.04	25	1	0	0
N068466	-1.31	1.65	0.04	23	1	0	0
N129150	-1.20	1.65	0.05	20	1	0	0
N289238	-0.69	2.77	0.07	24	5	1	0
N267256	-0.33	1.69	0.09	20	3	0	0
N131138	-0.04	1.60	0.11	21	4	0	0
N076034	0.16	1.56	0.13	22	5	0	0
N291267	0.49	1.48	0.18	19	6	0	0
N248399	0.87	1.41	0.27	15	7	0	0
N231222	1.34	1.23	0.47	11	9	0	0
N175476	1.58	1.10	0.62	9	11	0	1
N054456	1.87	0.96	0.84	6	13	0	6
N014401	2.12	0.79	1.09	5	21	0	1
M227221	2.23	0.77	1.23	3	17	0	9
M212203	2.33	0.74	1.37	2	16	0	0
M227218	2.48	0.69	1.65	1	15	0	5
M272293	2.52	0.59	1.73	1	27	0	0
M349360	2.82	1.09	2.39	2	17	1	1
M253292	3.00	0.32	2.76	0	25	0	0
M238199	3.00	0.35	2.77	0	18	0	0
M042016	3.00	0.98	2.77	1	17	1	0
M100128	3.00	0.35	2.77	0	19	0	8
M169159	3.32	3.52	3.47	9	9	11	0
M056047	3.50	0.64	4.18	0	19	1	8
M080067	3.69	0.78	5.52	0	14	2	4
M160093	3.91	1.37	8.24	1	20	10	1
M135098	4.11	0.87	12	0	21	9	0
M257251	4.36	3.15	18	3	5	8	2
M317305	4.94	3.16	**49**	3	6	12	0
M050058	5.24	1.41	**81**	0	7	17	2
M074062	5.52	1.46	**127**	0	7	23	0
M102094	6.00	3.43	**255**	3	6	20	0
M048058	6.28	1.61	**367**	0	4	27	0
M062013	6.63	2.99	**557**	1	3	15	0
M130138	7.12	1.67	**968**	0	1	18	0
M011013	7.70	4.61	**2076**	3	1	18	0
M045029	8.90	0.95	**10521**	0	0	16	0
M155096	8.96	3.46	**11216**	1	0	17	0
M320310	8.99	0.92	**11642**	0	0	21	0
M311295	9.05	0.91	**12270**	0	0	23	0
M086009	9.07	0.89	**12571**	0	0	28	0
M047051	9.19	3.42	**13998**	1	0	20	0
M173155	10.06	4.18	**22414**	2	0	33	0

the standard deviation of the prior distribution at 50 and reran the models. This had little effect on the likelihood ratios derived from the first dataset, and slightly decreased the likelihood ratios for the FBI/Noblis black box dataset, especially for μ values above 5.

A third assumption is the assumption of shrinkage for the values of σ for the normal distribution on the latent dimension. Shrinkage is a mechanism by which the values of σ act as mutually informative during the MCMC estimation process. This can benefit the parameter estimation because it can deal with extreme cases where the examiners were unanimous and the values of μ and σ are poorly constrained for unanimous pairs. However, we removed this constraint and replaced it with a generic and vague prior of a gamma distribution:

$\sigma_n \sim$ gamma(mode = 3.0, sd = 3.0). We re-ran the model fits from both datasets. For the FBI/Noblis black box dataset the likelihood ratios were slightly reduced, especially at higher values of μ, but these changes were small. For the Busey et al. [6] dataset, the likelihood ratios were also reduced, especially at higher values of μ. These effects are likely due to the fact that without shrinkage, some of the nonmated distribution variances are large and can extend into the range of 5–10 on the latent dimension where mated pairs tend to occur. There are lots of sensible reasons to include shrinkage to avoid these kinds of effects, and therefore we feel that this assumption is justified.

Given the above sensitivity analyses, we believe that the likelihood ratios are relatively insensitive to the choice of parameters and assumptions, and our current parameters provide the best estimates of the likelihood ratios for each image pair. The OSF.io site has a folder called *SensitivityAnalysisGraphs* that contains the graphs for these sensitivity analyses, and the source code contains flags to run these experiments. The readme.md document documents the different files found in that folder, as well as how to run the sensitivity analysis.

1.7. Interpreting likelihood ratios

A major contribution of the present work is that we have used human judgments to quantify the strength of evidence of fingerprint comparisons using the likelihood ratio. It may come as a surprise to some readers how low these values really are when compared to existing likelihood ratio calculations from other disciplines. Inspection of Table 1 and Table 2 provides a point of reference when interpreting the likelihood ratios. For example, consider the response distributions on the right side of Fig. 1. Although some examiners were willing to conclude Identification or Support for Common Source for this image pair, others were not willing or even Excluded. This leads to a likelihood ratio of 7.7, which is quite low relative to the astronomical levels observed in DNA, where values can easily exceed a billion. A statistically-literate jury member might interpret this likelihood ratio as follows (assuming the evidence was probative): However more times guilty than innocent the defendant was prior to hearing the fingerprint evidence, we would multiply that odds ratio by 7.7 to obtain a new odds ratio that reflects how much more guilty than innocent the defendant is. For example, if a jury member felt that the defendant was twice as likely to be guilty than innocent before the fingerprint testimony, the updated odds would be 15.4 times more likely to be guilty than innocent, assuming that the evidence was probative and the forensic practitioner seemed credible. Of course jury members can and do interpret these likelihood ratios in ways that do not align with the actual numerical value, typically underweighting the strength of the evidence [12,13].

Seen in this context, fingerprint evidence with low likelihood ratios can still produce a meaningful contribution to the case. When interpreted properly by the finder of fact, the relatively weak evidence in the image pair in Fig. 1 can be given appropriate weight but still be useful. In sum, the strength of the evidence provided by the image pair in Fig. 1 is not meaningless, but it also should not be given the same weight as an Identification on the image pair in Fig. 2 with its much higher likelihood ratio.

What determines the range of likelihood ratios that we see for typical casework? There are several factors.

First, consider the height of the thick blue curves in Fig. 8 and Fig. 11. There is a hump around 2.5 that corresponds to impressions that examiners agreed to compare, but many examiners ultimately reached an Inconclusive decision. Because the thick blue curve is normalized, this will reduce the height of the thick blue curve in the

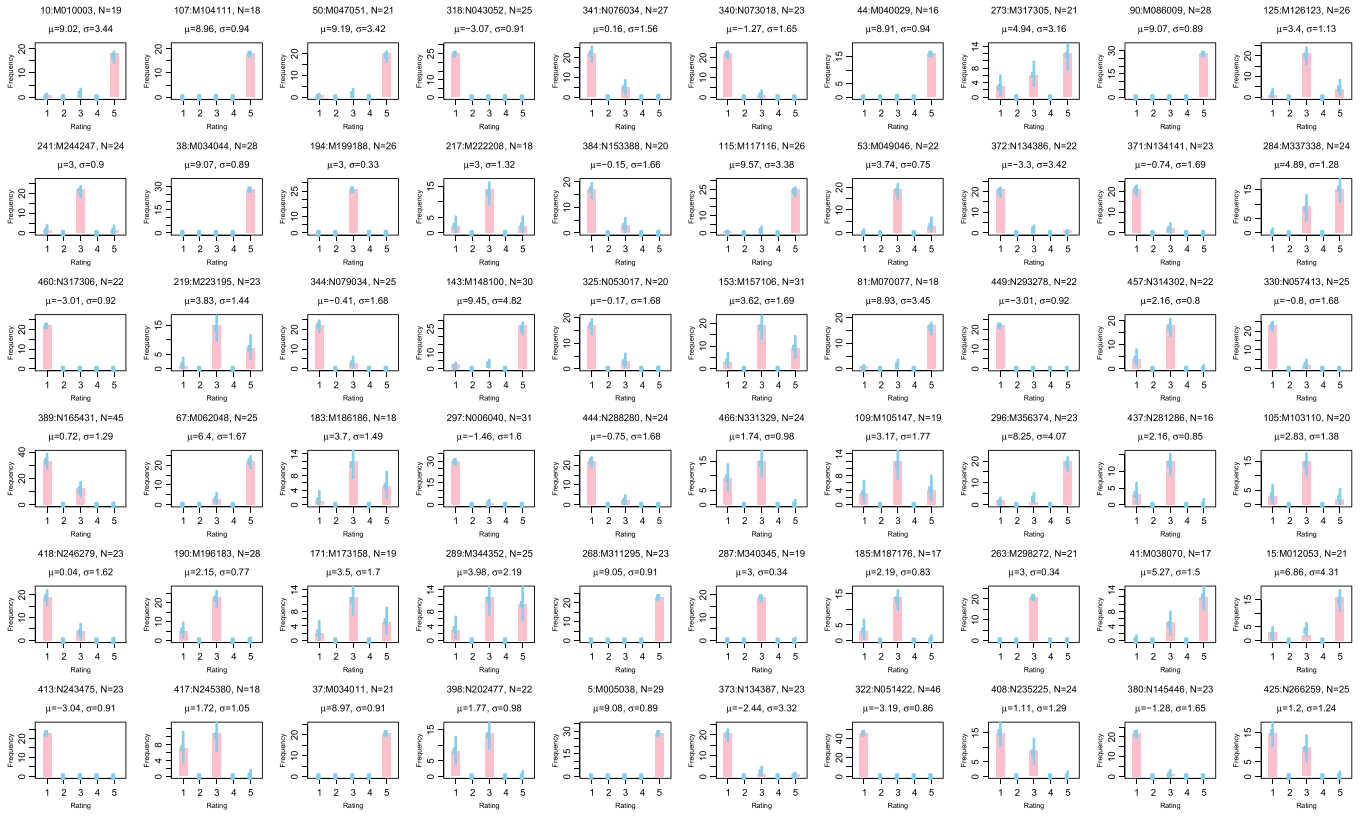

Fig. 10. Model fits to 60 randomly-selected pairs in the FBI/Noblis Black Box study [30]. The pink vertical bars provide the frequency of examiners who reach each conclusion, and the blue dots correspond to the predictions from the ordered probit model. See text for details.

range of about 5–10, which corresponds to typical casework. Essentially, the likelihood ratio values for higher-quality prints are being dragged down by the complex comparisons that cause disagreement or a majority of inconclusive decisions. Agencies could solve this by applying additional quality measures to complex impressions or applying a minimum standard to determine which impressions are classified as "of value". This would increase the likelihood ratio values for higher quality impressions, as image pairs with lower μvalues would no longer create the hump around 2.5 in the thick blue curves in Fig. 8 and Fig. 11. Defining what this standard would be is outside the scope of the present work and is an opportunity for future discussion.

Second, the likelihood ratios tend to depend on the behavior of the tail of the distributions. The original Black Box study was focused primarily on erroneous identification outcomes, where examiners conclude Individualization to nonmated pairs. It took tens of thousands of trials to observe just 6 of these errors, which demonstrates how difficult it is to estimate erroneous identification effects. However, the height of the thick red tail is not just determined by these few errors, but by all of the responses to all of the non-mated

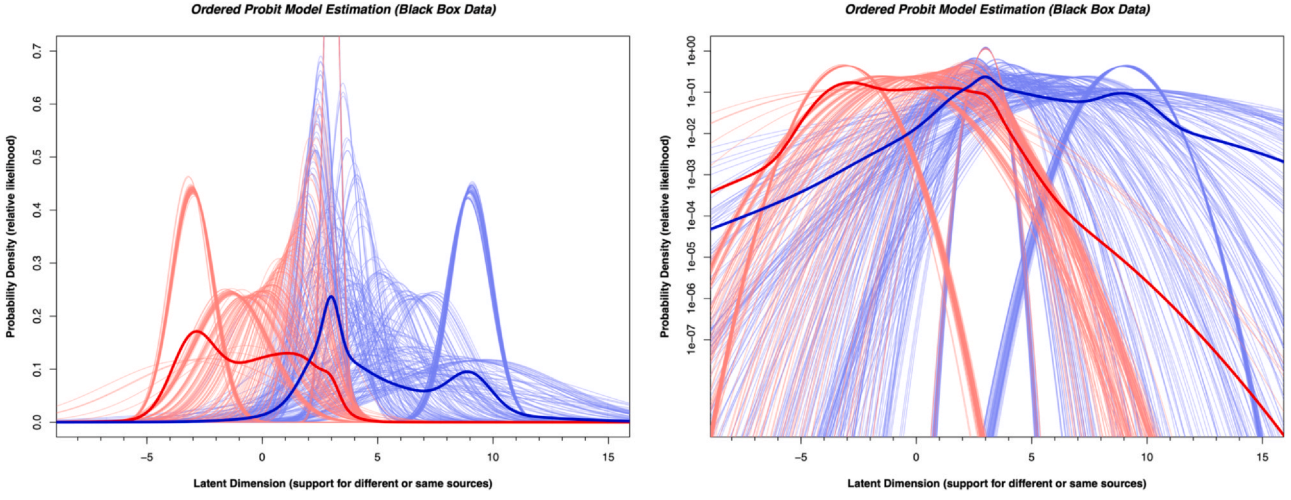

Fig. 11. Left panel: Relative likelihood of observing a given latent value for each mated (light blue curves) or nonmated (light red curves) comparison for the FBI/Noblis Black Box data [30]. The parameters for each normal distribution were derived from the ordered probit model fit to all three scales for each comparison. The thick red curve corresponds to the sum of light red curves, normalized to have an area of 1.0. It represents the relative likelihood of observing *any* nonmated comparison at each value of the latent axis. The thick blue curve represents the relative likelihood of observing *any* mated comparison at each value of the latent axis. Right panel: Same data plotted on a log(10) axis.

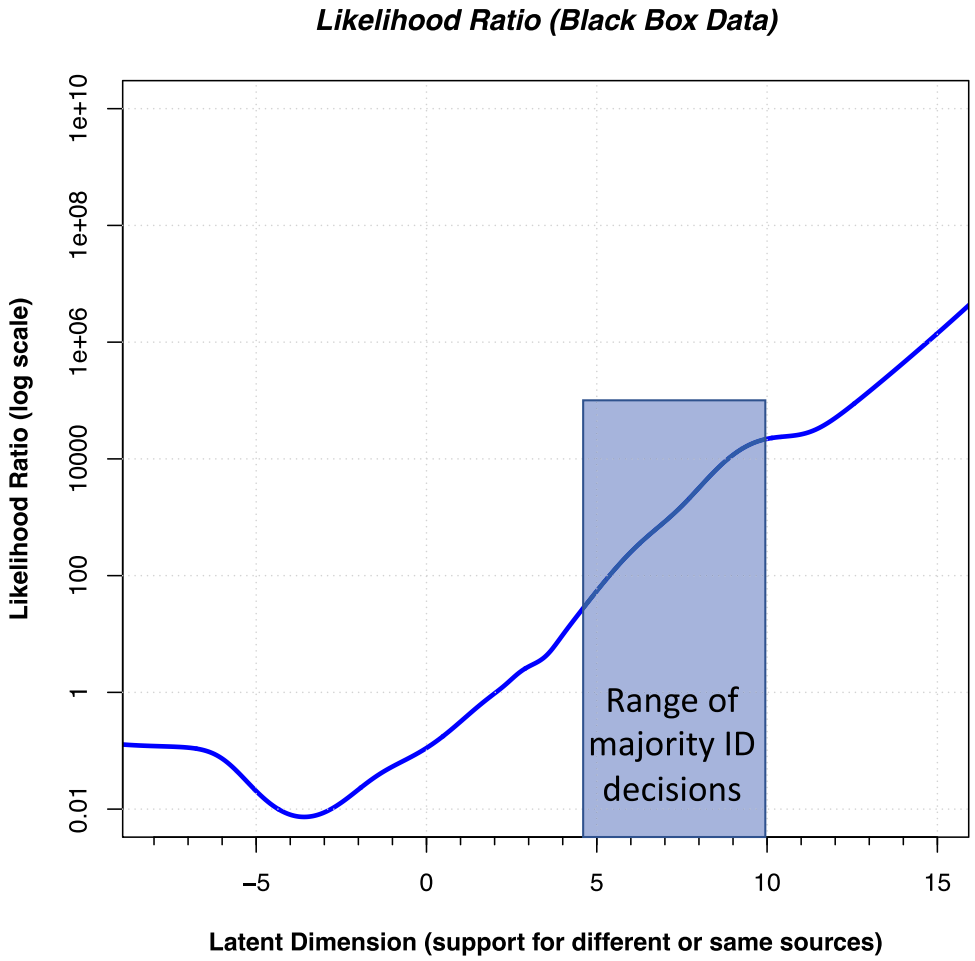

Fig. 12. Likelihood ratio values for different values along the latent axis for the FBI/Noblis Black Box data [30]. The y axis is plotted on a log(10) axis. The log of the likelihood ratio can be observed directly as the difference between the thick blue and thick red curves in the right panel of Fig. 11. The blue region illustrates the approximate range of image pairs with majority ID decisions.

image pairs. The utility of the ordered probit approach is that it translates categorical statements into an underlying distribution to explore the relation between the support for the same-source and different-sources propositions for all image pairs in the database. We explored various alternative distributional assumptions, but none produced larger likelihood ratios, and most distributions have heavier tails than the normal distribution.

Finally, the range of likelihood ratio values differs markedly from feature-based approaches such as Neumann et al. [15], which used triangulation of minutiae to derive features from which configurations could be compared. The measured likelihood ratios depend on the number of minutiae but are much higher than those reported in the present work. Although there is some debate about whether the likelihood ratio values reported in the paper depend on the choice of weighting functions (see comments at the end of their paper), typical LR values ranged from 10^5 to 10^{10} for mated pairs with between 5 and 10 minutiae (see Fig. 5 of Neumann et al. [15]). Why are these values so much higher than our likelihood ratio values that are based on human judgments? In the case of the present approach, we are not measuring the evidentiary strength of *fingerprints*, but instead we are measuring the evidentiary strength of *human judgments about fingerprints*. When a human testifies based on their observations and experience, our approach characterizes the evidentiary strength of that testimony. We are not arguing that our examiner consensus-based approach is superior to a feature-based approach; indeed, if a feature-based approach can augment or replace human judgments then it would obviously be preferred. However, the fact that algorithmic approaches demonstrate LRs of 10^5–10^{10} does not imply that human judgments therefore have similar evidentiary strength.

The range of typical likelihood ratios we observe may also align with our expectations based on the repeatability (intra-examiner) and reproducibility (inter-examiner) results reported by Ulery et al. [30,31]. While repeatability was high within examiners, reaching 90%, it was notably lower on image pairs noted as difficult. The authors attributed much of the variability to borderline cases, which the ordered probit model handles by having the latent distribution span across several decision categories. This will typically lead to lower μ values for many image pairs, creating the overlapping distributions seen in Fig. 8 and Fig. 11 between mated and nonmated pairs.

1.8. Implications for forensic science

How should we evaluate the range of likelihood ratios we observe? In our view, the range of typical casework values in the Busey et al. [6] dataset tends to fall in the range of 10–100,000 based on our subjective evaluation of the images in the folder *ImagesCombinedAndSorted* in the Supplementary Information and the fairly arbitrary criterion of receiving more Identification responses than all other responses combined (see bold numbers in Table 1). Likelihood ratios were similar in the Black Box error rate study data, ranging from 50 to 20,000. These values are quite modest relative to verbal scales produced by DNA analysts. For example, SWGDAM [26]

has a verbal equivalent scale in which the range of 100–10,000 is listed as "Moderate support for inclusion". The Association of Forensic Science Providers have a verbal equivalence scale [3] that lists a range of 10–100 as Moderate Support, 100–1000 as Moderately Strong Support, 1000–10,000 as Strong Support, and 10,000–1000,000 as Very Strong Support. These ranges are very different than the Extremely Strong Support for Common Source language that has been proposed for an expanded fingerprint scale [19] which would require a likelihood ratio of over a million on the Association of Forensic Science Providers scale. It is unlikely that Extremely Strong Support for Common Source would be used synonymously as Identification, and in Busey et al. [6] examiners reserved used this statement less often than they used Identification. However, this statement might be viewed as a drop-in replacement for Identification by some examiners (which by itself might be overinterpreted by laypersons, see Swofford and Cino [27]). Given the range of likelihood ratios observed in the present work, the term Identification may only be appropriate for very clear impressions, which also provides an argument for expanded conclusion scales that include phrases such as Support for Common Source.

The ordered probit model likely underestimates the true evidentiary value of very clear impressions that receive unanimous conclusions. Consider the impression in Fig. 2, which was unanimous across all three scales. The ordered probit model will try to accommodate unanimity by making μ as large as possible, and in the Bayesian case, μ is constrained only by the prior placed on μ. Without this constraint, the model would essentially make μ infinite for unanimous comparisons, producing infinite likelihood ratio values. The true likelihood ratio of such an image pair is likely much larger than the ~110,000 value that is estimated by the model, but this image pair represents such high image quality and quantity that the quantification of the strength of the evidence is almost unnecessary because the amount of agreement is apparent to even the untrained eye. However, this is not true for the bulk of casework-like impressions where likelihood ratios are in the range of 50–20,000. For these casework-like image pairs, the model is well-behaved in its estimate of $\mu$and therefore the likelihood ratio is an accurate estimate of the strength of support for the same source and different sources propositions.

1.9. Comparing disciplines

Our use of the ordered probit model to estimate likelihood ratios for individual image pairs readily extends to any forensic discipline for which error rate or black box testing is available. This model answers calls for a unified evidence scale across disciplines [18]. A nice feature of our approach is that while many forensic disciplines tend to use similar language (i.e. Identification is often the highest category of responses), we anticipate that the likelihood ratios obtained across disciplines will vary widely, which may mirror the lay understanding of the relative strengths of different disciplines (Thompson & Newman, 2015). By computing the range of likelihood ratios for typical casework-like comparisons in a variety of forensic disciplines, the ordered probit model could help jurors and prosecutors appropriately weigh the evidence from different forensic disciplines. The use of the ordered probit model readily accommodates laboratories that use different conclusion scales (or even traditional vs expanded). The model can also be applied where agencies use subjective likelihood ratios [2], because the distribution of subjective likelihood values from ground-truthed images can be directly modeled with a latent normal distribution without the need for thresholds.

The likelihood ratio values are independent of the number of mated and nonmated pairs in the database, because each set is normalized separately. One advantage of likelihood ratios is that they are independent of the prior probability of a mated pair (i.e. how good your detectives are or how big your database is). A finder of fact can simply form their own prior odds and multiply these by the likelihood ratio to get an updated posterior odds value. The likelihood ratio is also independent of the size of the conclusion scale or the wording used to construct the conclusion scale, as long as the ordered probit model is constructed properly. Thus, the current approach can provide quantitative measures of the strength of evidence across forensic disciplines, laboratories, and reporting styles.

1.10. Applications to casework

We believe that the current approach might also work for active casework comparisons to provide a likelihood ratio for each comparison. Suppose we knew the likelihood ratios for six fiduciary image pairs that are derived from distributions of examiner responses in an error rate study (and the present work provides one source for these likelihood ratios). An examiner could individually examine each fiduciary image pair to determine how much support each image pair provided for the same source proposition. To do this, they would assess the relation between the two images as if they were conducting a comparison, but instead of reaching a decision they would simply note the amount of support for the same source proposition. This is a subjective process much like the decision in casework, and it is based on a comparison between the physical features of the two impressions. These judgments are subject to the same inter-examiner variability and biases that occur in casework, but these can be minimized as explained below.

When a casework image pair is compared, the examiner would determine how much support this image pair offered for the same source proposition, *and this value could be compared against each of the six fiduciary image pairs*. By ranking the casework value against the values provided by the fiduciary image pairs, we can determine a likelihood ratio value for the casework image pair by reference to the known likelihood ratios from the fiduciary image pairs.

There are a variety of ways that we might accomplish this comparison between casework and fiduciary image pairs. We are exploring a method developed by Saaty [23] that allows us to place the current casework image pair relative to the six fiduciary comparisons, each of which has a known location along the latent dimension in Fig. 9. This comparison involves choosing from a selection of statements such as "The casework image pair provides Definitively more support for the same source proposition than the current fiduciary image pair." The range of modifiers include Demonstrably, Moderately, Slightly, and Equal Support, and the statements are reversed to allow for the current fiduciary image pair to provide more support than the casework image pair for the same source proposition. These statements are then converted to a metric axis using an Eigenvector decomposition as detailed in [23], which locates all six fiduciary image pairs and the current casework image pair along this metric axis. The fiduciary values along the metric axis can be regressed against their known likelihood ratio values, and the regression equation provides an estimate of the likelihood ratio for the casework image pair through its metric value.

These comparative judgments would be subject to the same inter-examiner variability and biases found with traditional decisions. The greatest concern might come from an examiner who felt that a particular casework pair provided much more support for the common source proposition than their peers feel is warranted. This might produce an unjustifiably high likelihood ratio by the proposed method. There are three elements that ameliorate this concern. First, the relation between the estimated metric values from the subjective judgments and the calculated likelihood ratios might not correlate for the six fiduciary image pairs, which would serve as a warning sign that the Eigenvalue solution is inaccurate for this particular examiner. Second, labs could conduct technical review of

two independently-estimated likelihood ratios to determine whether they are within some acceptable range. Finally, image pairs with known likelihood ratios could be inserted into the casework workflow to identify whether the examiner-estimated likelihood ratios are within some acceptable range of the likelihood ratios calculated from error rate studies. Proficiency testing would be done in similar manner. Thus, casework would proceed in a similar manner as currently practiced in the United States, with the exception that examiners would offer observations instead of conclusions, and those observations would be in the form of likelihood ratios that directly provide the strength of support for the same- and different-source propositions.

The likelihood ratios could also be related directly to objective measures of feature quality, quantity, and rarity, such as those provided by LQMetric (Ulery et al. [34]) and FRStat [28]. This would provide an association between physical measures and the examiner-based likelihood ratios. The challenge with this approach is that it requires a large number of image pairs to measure the association. The data from Busey et al. [6] has too few likelihood ratios from majority-ID image pairs, and the Ulery et al. [30] images are not publicly available. However, future work could explore various image-based analyses on image pairs that are both available and have error rate data.

Should this proposed approach prove reliable, it would provide the means to quantitatively measure the strength of fingerprint evidence, and the methods could be extended to operational casework. Once validated, the techniques above could be used in any forensic discipline for which error rate studies are available.

CRediT authorship contribution statement

Thomas Busey: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Project administration. **Meredith Coon**: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful to John Kruschke for sharing the ordered probit model code.

References

[1] N.R.C. o t N. Academies, Strengthening Forensic Science in the United States: A Path Forward, National Academy Press, 2009.
[2] Aitken, C., Barrett, A., Berger, C., Biedermann, A., Champod, C., Hicks, T.,. McKenna, L. (2015). ENFSI guideline for evaluative reporting in forensic science.
[3] Assoc Forensic Sci Providers, Standards for the formulation of evaluative forensic science expert opinion, Justice 49 (3) (2009) 161–164, https://doi.org/10.1016/j.scijus.2009.07.004
[4] J. Buckleton, B. Robertson, J. Curran, C. Berger, D. Taylor, J.-A. Bright, S. Pugh, A review of likelihood ratios in forensic science based on a critique of Stiffelman "No longer the Gold standard: Probabilistic genotyping is changing the nature of DNA evidence in criminal trials", Forensic Sci. Int. 310 (2020) 110251.
[5] Busey, T., & Klutzke, M. (2022). Calibrating the Perceived Strength of Evidence of Forensic Testimony Statements. Science and Justice.
[6] T. Busey, M. Klutzke, A. Nuzzi, J. Vanderkolk, Validating strength-of-support conclusion scales for fingerprint, footwear, and toolmark impressions, J. Forensic Sci. 67 (3) (2022) 936–954.
[7] T. Busey, N. Heise, R.A. Hicklin, B.T. Ulery, J. Buscaglia, Characterizing missed identifications and errors in latent fingerprint comparisons using eye-tracking data, Plos One 16 (5) (2021).
[8] K.E. Carter, M.D. Vogelsang, J. Vanderkolk, T. Busey, The utility of expanded conclusion scales during latent print examinations, J. Forensic Sci. 65 (4) (2020) 1141–1154.
[9] I.W. Evett, Towards a uniform framework for reporting opinions in forensic science casework, Sci. Justice 38 (3) (1998) 198–202 doi:Doi 10.1016/S1355-0306(98)72105-7.
[10] IAI. (2010). IAI Resolution 2010–18. In (Vol. 2010): International Association for Identification.
[11] Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.
[12] K.A. Martire, R.I. Kemp, B.R. Newell, The psychology of interpreting expert evaluative opinions, Aust. J. Forensic Sci. 45 (3) (2013) 305–314, https://doi.org/10.1080/00450618.2013.784361
[13] K.A. Martire, R.I. Kemp, M. Sayle, B.R. Newell, On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect, Forensic Sci. Int. 240 (2014) 61–68, https://doi.org/10.1016/j.forsciint.2014.04.005
[14] L. Mickes, J.T. Wixted, P.E. Wais, A direct test of the unequal-variance signal detection model of recognition memory, Psychon. Bull. Rev. 14 (5) (2007) 858–865 doi:Doi 10.3758/Bf03194112.
[15] C. Neumann, I.W. Evett, J. Skerrett, Quantifying the weight of evidence from a forensic fingerprint comparison: a new paradigm, J. R. Stat. Soc.: Ser. A (Stat. Soc.) 175 (2) (2012) 371–415.
[16] C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, A. Bromage-Griffiths, Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae, J. Forensic Sci. 52 (1) (2007) 54–64.
[17] C. Neumann, C. Champod, R. Puch-Solis, N. Egli, A. Anthonioz, D. Meuwly, A. Bromage-Griffiths, Computation of likelihood ratios in fingerprint identification for configurations of three minutiae, J. Forensic Sci. 51 (6) (2006) 1255–1266.
[18] A. Nordgaard, R. Ansell, W. Drotz, L. Jaeger, Scale of conclusions for the value of evidence, Law, Probab. Risk 11 (1) (2012) 1–24, https://doi.org/10.1093/lpr/mgr020
[19] OSAC, F.R.S. (2018). Standard for Friction Ridge Examination Conclusions [DRAFT DOCUMENT].:
[20] PCAST. (2016). Ensuring Scientific Validity of Feature-Comparison Methods. Retrieved from.
[21] Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Paper presented at the Proceedings of the 3rd international workshop on distributed statistical computing.
[22] Plummer, M. (2012). JAGS Version 3.3. 0 user manual. In: Lyon, France.
[23] T.L. Saaty, A scaling method for priorities in hierarchical structures, J. Math. Psychol. 15 (3) (1977) 234–281.
[24] B.A. Spellman, Communicating forensic evidence: lessons from psychological science, Seton Hall. L. Rev. 48 (2017) 827.
[25] Swanson, C.L. (2020, November 20, 2020). [USACIL DFSC conclusion scale].
[26] SWGDAM. (2018). Recommendations of the swgdam Ad hoc working group on genotyping results reported as likelihood ratios.
[27] H.J. Swofford, J.G. Cino, Lay understanding of "identification": how jurors interpret forensic identification testimony, J. Forensic Identif. 68 (1) (2017) 29–41.
[28] H.J. Swofford, A.J. Koertner, F. Zemp, M. Ausdemore, A. Liu, M.J. Salyards, A method for the statistical interpretation of friction ridge skin impression evidence: method development and validation, Forensic Sci. Int. 287 (2018) 113–126.
[29] W.C. Thompson, E.L. Schumann, Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy, In Expert Evidence and Scientific Proof in Criminal Trials, Routledge, 2017, pp. 371–391.
[30] Ulery, B.T., Hicklin, R.A., Buscaglia, J., & Roberts, M.A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. Proceedings of the National Academy of Sciences of the United States of America, 108(19), 7733–7738. 10.1073/Pnas.1018707108.
[31] Ulery, B.T., Hicklin, R.A., Buscaglia, J., & Roberts, M.A. (2012). Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners. PloS One, 7(3), 1–12. doi:ARTN e3280010.1371/journal.pone.0032800.
[32] B.T. Ulery, R.A. Hicklin, M.A. Roberts, J. Buscaglia, Changes in latent fingerprint examiners' markup between analysis and comparison, Forensic Sci. Int. 247 (2015) 54–61, https://doi.org/10.1016/j.forsciint.2014.11.021
[33] B.T. Ulery, R.A. Hicklin, M.A. Roberts, J. Buscaglia, Interexaminer variation of minutia markup on latent fingerprints, Forensic Sci. Int. 264 (2016) 89–99, https://doi.org/10.1016/j.forsciint.2016.03.014
[34] B.T. Ulery, R.A. Hicklin, M.A. Roberts, J. Buscaglia, Factors associated with latent fingerprint exclusion determinations, Forensic Sci. Int. 275 (2017) 65–75, https://doi.org/10.1016/j.forsciint.2017.02.011